Exploring Bias

Article | No 78

EUROPEAN FUTURES

Exploring Bias: Comparing Approaches for Collecting Twitter Data

Author(s): Clare Llewellyn, Laura Cram

Permalink: http://www.europeanfutures.ed.ac.uk/article-2806

Publication: 10 February 2016

Article text:

*As part of the Imagine Europe project, **Clare Llewellyn** and **Laura Cram** explore the different kinds of Twitter data sets they have collected on the UK's EU referendum debate and the insights that each set's distinct characteristics can provide.*

In our Imagine Europe project, we are tracking the UK's EU referendum debate to explore the various ways in which the public imagines the European Union. We are using Twitter to map trends in response to emerging events. This analysis allows us to gain a more nuanced understanding of those who are motivated to comment on UK-EU-related topics. See our Twitter demo for interactive visualisations of the data.

We have collected Twitter data on the referendum debate for the past five months. We are using three methods for collecting data from Twitter: 1) Using hashtags chosen by an expert panel as search queries; 2) Collecting a random sample without specified search terms and extracting referendum-appropriate data automatically; 3) Collecting from the three official campaign groups @vote leave, @LeaveEUOfficial and @StrongerIn.

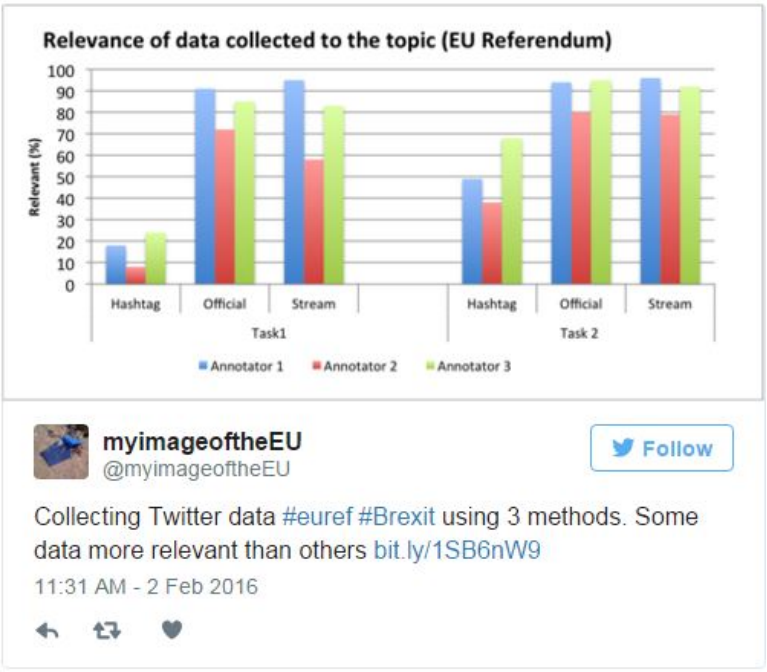
Each method of collection influences what data will be collected and therefore each data set has certain biases. The hashtag and random stream sample sets are heavily influenced by the terms used for data collection. Those terms differ greatly when automatically extracted (the random stream set) or chosen by experts (the hashtag set).

The expert method is designed to follow a wider variety of terms that the experts expect will become discussion topics over the longer-term referendum debate, whereas the automatic method extracts data using terms which are commonly associated with known referendum-specific terms. Examining the three different sets allows us to contrast what is being collected and gives us the ability to have a broader understanding of public and elite opinion. In particular, we are examining how topics differ between these data sets and how they influence each other.

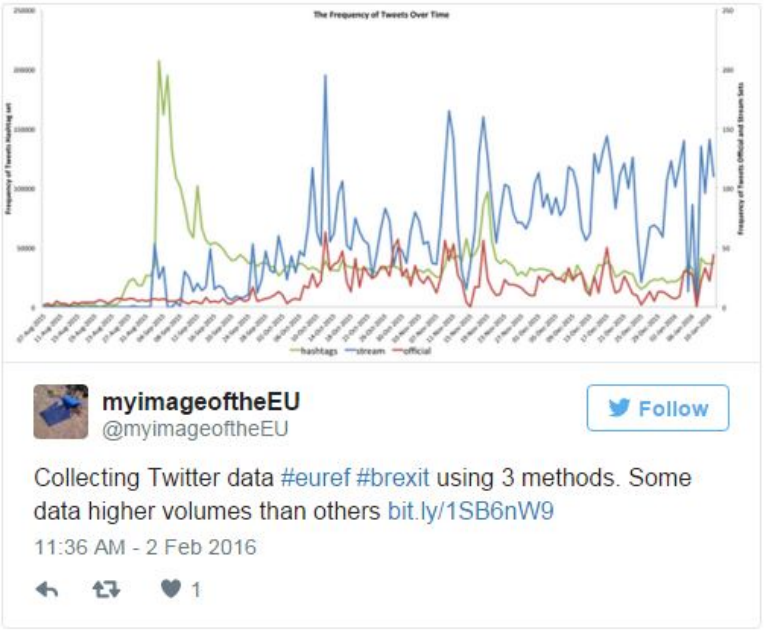
The hashtag set is the largest by a considerable amount. During the five-month period, the set collected using hashtags contained 5,556,027 tweets. The set extracted from the random stream has 8,777 tweets and the official campaigns 2,606 tweets. To determine how relevant the data collection is to the debate, we extracted 100 tweets from each set and asked three annotators to consider the

relevance of each tweet in two ways: 1) whether it is directly relevant to the UK-EU referendum debate, and 2) whether it is about a topic that would likely influence voter opinion.

We found that the data from the official campaign groups and data automatically extracted from the random stream are more relevant to the topic than the data gathered using hashtags. The hashtag set has a low relevance score for 'directly relevant to the referendum debate' but this rises significantly when the topics that will influence the debate are considered.

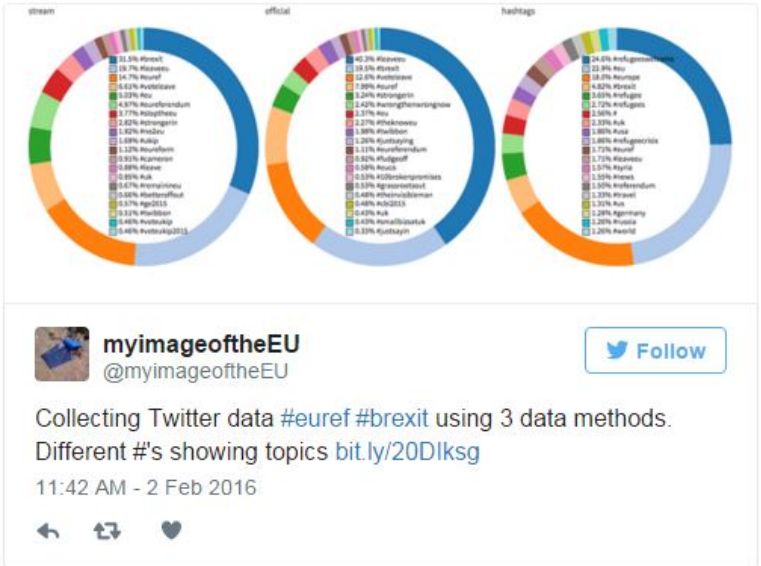


This was as we expected. The differences can be explained as follows. The official campaign set contains the information from the campaign groups which are publishing tweets in order to influence the debate. This gives us a small, very specific, very opinion-driven set. The random stream set gives a set of data from the wider public, but only tweets that contain terms that are closely related to the debate, therefore providing a very topic-specific set. The hashtag-gathered set is a much larger set, collected using a wider variety of terms. It contains more non-relevant information but also covers the topics likely to influence voters not identified in the other sets.



When we look at tweeting frequency over time, we find that all of the collection strategies are picking up increases in data volumes on the same dates. This takes place when there are events that prompt both the campaigns and general Twitter users to tweet. For example, there is a peak in data collected on 12 October 2015, when the Britain Stronger in Europe campaign was launched. A further peak coincides with David Cameron's speech at Chatham House setting out his case for EU reform and his letter to European Council President Donald Tusk on 10 November 2015.

We did find that much more data was collected by the hashtag method in early September. On further inspection this data relates to refugees and migrants. This shows that the campaign groups are not talking about the refugee crisis or related migration issues. It is not being directly related to the UK-EU referendum debate, but instead it is being widely discussed.



Analysing the frequency of commonly used hashtags gives an indication of topics discussed in each of the datasets. Much of the discussion in the tweets from the official campaign and the random stream data are directly related to the UK-EU referendum. This is echoed by the hashtags #brexit, #leaveeu, #voteleave and #euref being the top four most frequent in both collections.

Hashtags with a pro-Leave sentiment appear more frequently in all three of our data sets. We do not see any pro-Remain hashtags appearing in the hashtag-gathered set, and only #strongerin and #remainineu in the random stream set. We have a very small number of pro-Remain hashtags in the official campaign data.

Since we are collecting from the three campaign groups and only one is pro-Remain, we would expect a lower level of pro-Remain hashtags in the official set – but not as low as we are seeing. This suggests that either pro-Remain supporters don't use hashtags, use them in unexpected ways or there is a strong pro-Leave sentiment on Twitter.

We also see another phenomenon within the data – where hashtags are used to draw attention to specific themes. Within the official stream, certain hashtags have been heavily used by the two pro-Leave campaigns. For example, @LeaveEUOfficial launched #theknoweu, #justsaying, #fudgeoff and #twibbon and @vote_leave launched #wrongthenwrongnow and #theinvisableman. We can see that #twibbon also appears in the random stream data set and therefore has cross-pollinated and is being discussed by the wider public. The @StrongerIn campaign does not seem to be using hashtags to the same extent and rarely uses any beyond #strongerin. It is possible that the lack of use of hashtags by the @StrongerIn campaign means that their supporters are not using hashtags as well. This is something we will need to investigate further.

In the hashtag-gathered data, many of the top hashtags indicate a focus on the topic of refugees (#refugeeswelcome, #refugee, #refugeecrisis) and in discussing

particular countries (#uk, #usa, #syria, #germany). In the random stream data, we also see a discussion of the referendum-specific terms #brexit and #leaveeu, but very little occurrence of the #strongerin hashtag.

Our #ImagineEurope project is part of the Economic and Social Research Council's The UK in a Changing Europe programme. Look out for our regular updates as the project tracks developments in the debate on the UK's membership of the EU and follow us on Twitter @myimageoftheEU for more information on this and other projects.

Laura Cram is Senior Fellow, The UK in a Changing Europe, investigating The European Union in the Public Imagination: Maximising the Impact of Transdisciplinary Insights *(ESRC/ES/N003985/1).*

This article was originally published on the ImagineEurope Storify.

Author information:

Clare Llewellyn
The University of Edinburgh

Clare Llewellyn is PhD Candidate in Informatics at the University of Edinburgh and Research Assistant in the European Union in the Public Imagination project. Her research focuses on user-generated content on the Internet.

Laura Cram
The University of Edinburgh

Prof Laura Cram is Professor of European Politics at the University of Edinburgh; Senior Fellow, The UK in a Changing Europe; and Academic Editor of *European Futures*. Her research areas include European public policy, European identity and the neuropolitics of public policy and identity.

Additional information:
Please note that this article represents the view of the author(s) alone and not European Futures, the Edinburgh Europa Institute nor the University of Edinburgh.